Transactions on Ecology and the Environment vol 19, © 1998 WIT Press, www.witpress.com, ISSN 1743-3541

A reinforcement learning model for the operation of conjunctive use schemes

F. J.-C. Bouchart & H. Chkam
Department of Civil & Offshore Engineering, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom
EMail: bouchart@civ.hw.ac.uk

Abstract

A Reinforcement Learning (RL) model is proposed to identify operational strategies for conjunctive use schemes. This model is based on neuron-like adaptive elements that learn on-line to avoid system failures. The strength of this approach is that the resulting operational strategy for the water supply scheme reflects the need to respond and adapt to discrete failure events. A second advantage of the RL methodology is the inherent ability of control elements to operate in a distributed manner. By responding to local state inputs combined with a combination of local and global performance signals, the individual control elements are capable of operating effectively with only a limited set of state variables. The implication of such localised control is the avoidance of the curse of dimensionality commonly exhibited in other methodologies. Application of the RL model is demonstrated using the Burncrooks reservoir complex in Scotland. The model learns to effectively avoid failures, resulting in improved operational reliability.

Introduction

The operation of any water supply system requires a sequence of decisions to be made, where the resulting performance of the system is derived from the cumulative effect of these decisions rather than any one single decision in isolation. This distributed impact of individual decisions becomes even more pronounced when developing operational strategies for conjunctive use schemes, where multiple reservoirs (or other water sources) are operated in a co-ordinated manner. The challenge is to develop crisp control rules for the system which are to be subsequently adhered to at discrete decision points, while only their cumulative impact can be assessed.

A further complicating factor is the existence of critical performance indicators related to customer service which are derived from discrete events when the supply system fails to provide an adequate level of service, as defined by water quantity and quality characteristics. These failure events provide key feedback on the performance of the system, thereby indicating the need for modifications to the operational policies employed. Dynamic Programming and Linear Programming formulations employ explicit functions for the contribution of each operational decision towards the final performance objective, thereby providing guidance for the operational modifications necessary to avoid system failures and reductions in levels of service[1]. The challenge is that individual contributions towards achieving adequate levels of service cannot be identified without the use of artificial heuristic rules. The assessment of customer service is a global indicator of performance of the system up to the time of failure. Therefore, this failure cannot be disaggregated into the individual time steps leading to the failure event. Methodologies capable adapting to individual complaints from consumers must incorporate a framework to address the discrete, and often delayed, feedback from customers.

A Reinforcement Learning (RL) model is proposed that adapts to individual failure events without the need to develop an explicit objective function which disaggregates the performance of the system at discrete time steps. The remainder of the paper presents the RL model formulation before demonstrating the application of this model using the Burncrooks reservoir complex in Scotland.

Reinforcement Learning Model

The Reinforcement Learning (RL) model employed in this study is based on the work of Barto et al. [2]. The model is composed of neuron-like adaptive elements capable of solving complex control problems. These elements identify appropriate control strategies through an on-line learning process, depicted in Figure 1. The state (or condition) of the physical system, $\mathbf{z} = \{z_i\}$, at a given point in time is observed, and given this state the control element initiates a control decision, y. The response of the physical system to this control signal is then evaluated and a

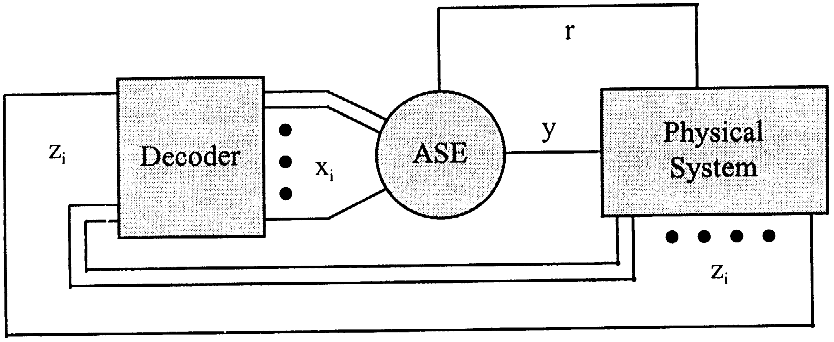


Figure 1: Reinforcement Learning control element

measure of performance obtained. Feedback to the control element is provided in the form of a reinforcement signal, r. This reinforcement signal rewards control decisions leading to the ideal system performance, while discouraging decisions leading to system failures. In this way the control element modifies its behaviour to increase the likelihood of rewards while avoiding punishment. These modifications are achieved without the aid of an external teacher ('teacher' in this case is used in the context of Neural Network theory and supervised learning), and the control element must learn to adapt through experience. The task of the control element is to learn to select sequences of actions that produce the greatest cumulative reward.

The control element consists of a decoder which maps the state vector **z** onto a discretised representation of the state space. The state space for the physical system, defined by the *n* state variables, is divided into M *n*-dimensional 'boxes' and each box *i* assigned a variable x_i. A system state lying within box *i* is then represented as $x_i = 1$ and $x_j = 0$ for all $j \neq i$. Individual control strategies, defined by the value of y, are then developed for each of the boxes within the discrete state space representation. The second part of the control element is a neuron-like element referred to as the Adaptive Search Element (ASE). The ASE evaluates the weighted sum of the input signal **x** and emits an output signal y which corresponds to the appropriate control decision. The learning process is embodied within the weights, $\mathbf{w} = \{w_i\}$, of the ASE. Mathematically, the control signal y at time *t* is defined as,

$$y(t) = \psi\left[\sum_{i=1}^{M} w_i(t)x_i(t) + noise(t)\right] \qquad (1)$$

The *noise(t)* term in eqn (1) is a random variable following a Gaussian distribution with a mean of zero and a variance σ^2. $\psi[\bullet]$ is a threshold function used to distinguish between control decisions. In the case of an ON/OFF control system, $\psi[\bullet]$ can be defined such that,

$$\psi[u] = \begin{cases} +1 & if\ u \geq 0 & [ON] \\ -1 & if\ u < 0 & [OFF] \end{cases} \tag{2}$$

The behaviour of the ASE adapts to its environment by modifying the weights vector **w**. In the case of the ON/OFF controller where in the system is in box *i*, positive values of w_i increase the likelihood of an ON signal (y = +1), while a negative value of w_i increases the likelihood of an OFF signal (y = -1). The *noise(t)* term provides a random behaviour component which facilitates the exploration of the complete decision space. Prior to learning, the weights vector **w** is initialised using random values. Individual weights w_i are then subsequently modified during the learning process. The revised value for the weight w_i at time t+1 is calculated using,

$$w_i(t+1) = w_i(t) + \alpha r(t) e_i(t) \tag{3}$$

The α coefficient defines the rate of change of the weights. Because of the presence of the reinforcement signal r(t) in the second term of eqn (3), the weights of the ASE are only modified when a reinforcement signal is emitted. If reinforcement is limited to a negative signal r(t) = -1 corresponding to a failure of the system at time *t*, then modifications to the weights only occur when a failure occurs. The $e_i(t)$ term in eqn (3) is the *eligibility* trace of box *i*. When a failure occurs in the system at time *t*, this failure is the result of the cumulative effect of a sequence of decisions and not simply the result of the last decision made, y(t). Therefore, a mechanism must be developed that allows the blame for the failure to be allocated amongst the sequence of decisions which led to this failure[3]. The mechanism adopted in this study is one where more recent decisions (alternatively, state "boxes" which have been traversed most recently) are allocated more blame than decisions made in the more distant past. The eligibility trace is then a function that decays over time, with discontinuous steps occurring whenever box *i* is entered. Furthermore, the eligibility trace reflects the predominant decision made when in box *i* by defining the discontinuous steps as a function of the output signal y(t) of the ASE. This last component of the definition of the eligibility trace ensures that the weights are modified to reduce the likelihood of the predominant decision (when in box *i*), which may have contributed to the failure event. A more detailed presentation of this mechanism can be found in [2].

One important modification to the RL model developed for the current study is the expansion of the control element to allow for control settings other than simple ON/OFF switches. The output signal y(t) of the ASE element is replaced by a vector signal **y**(t) of size K, thereby providing K+1 control settings for the system. Each $y_k(t)$ value is calculated using the function defined by eqn (2) and the following equation replacing eqn (1),

$$
\begin{aligned}
& y_1(t) = \psi\left[\sum_{i=1}^{M} w_{i,1}(t)x_i(t) + noise_1(t) + bias_1\right] \\
& \textit{if } y_1(t) \ge 0 \\
& \quad \textit{then} \quad y_2(t) = \psi\left[\sum_{i=1}^{M} w_{i,2}(t)x_i(t) + noise_2(t) + bias_2\right] \\
& \quad \textit{otherwise } y_2(t) = -1 \\
& \textit{and generally,} \\
& \textit{if } y_{k-1}(t) \ge 0 \\
& \quad \textit{then} \quad y_k(t) = \psi\left[\sum_{i=1}^{M} w_{i,k}(t)x_i(t) + noise_k(t) + bias_k\right] \\
& \quad \textit{otherwise } y_k(t) = -1
\end{aligned}
\qquad (4)
$$

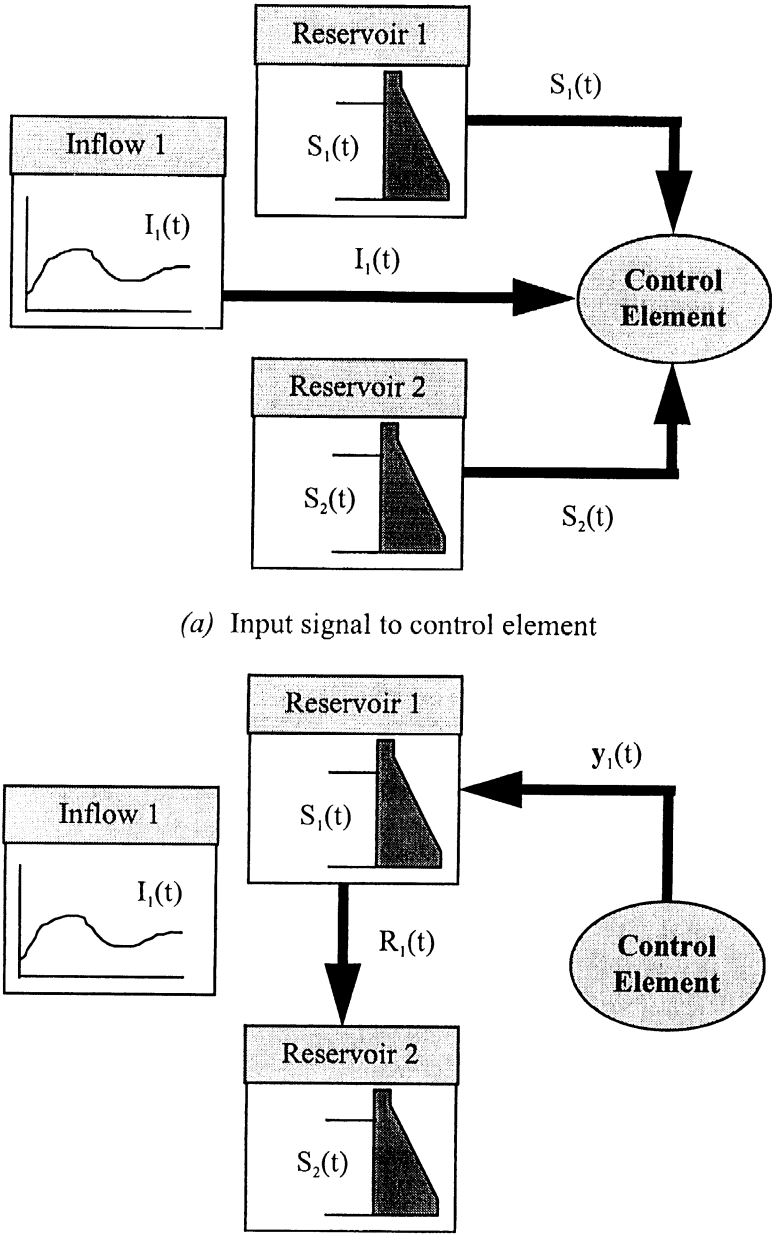
The eqn (4) represents a “cascade” function where a positive value of $y_{k-1}(t)$ implies a higher setting has been selected, using an ordering of settings where the lowest setting is **y(t)** = [-1,-1,-1] and the highest is **y(t)** = [+1,+1,+1] for the case where K = 3. It should be noted that separate weights are used for the different “levels” of this cascade, and a $bias_k$ term as been added to provide greater flexibility to the formulation. These bias terms can be set such that there is a greater likelihood of obtaining a “low” **y(t)** output signal. This would be useful for example in the case of a pumped system where it is best to keep the pump settings as low as possible. Only if training indicates that the pumps should be set at a higher level would a higher setting be sought. The implication is that the modified algorithm biases the search for a control strategy that employs low rather than high pump settings. This is important in the view that the RL approach is not an optimisation algorithm, and therefore, the search (learning) terminates when system failures no longer occur.

For the multiple reservoir operation problem which is the focus of the current study, a control element can be defined to control water transfers between subsystems of the whole multiple reservoir system. Consider for example a reduced representation of a water supply system, depicted in Figure 2, and consisting of two reservoirs and the inflow to Reservoir 1. A control element can be developed that takes as its input the storage levels of the two reservoirs and the inflow to Reservoir 1, and returns a signal corresponding to the water transfer to be made between Reservoir 1 and Reservoir 2. This element can then be trained to avoid failure events using either on-line real-time learning, or a simulation of the real system to avoid the problems of real-time operation using an untrained controller. It should be noted that the second option requires enough information about the physical system to establish a realistic simulation of the system.

As already stated, the control element for the two reservoir system is trained to avoid failures. These failure events can include spillage of water when excess reservoir capacity is available, storage levels in either reservoirs dropping below specified minimum levels, the inability to meet conpensation releases, etc. In addition, it was stipulated at the beginning of the study that it might be possible to incorporate failure events that were outside the confines of the modelled system. For example, the system depicted in Figure 2 could be part of a much larger system containing additional sources as well as demand nodes. Failure to supply one of these demand nodes could be a trigger to changing the release strategy for Reservoir 1 if such a change could eventually affect the performance of the demand node. The advantage of this approach is the ability to avoid the *curse of dimensionality* present in such approaches as dynamic programming. If a control element can be developed to respond to global failure events without having to model the state space of the complete system, then it should operate effectively at a local level based on local input, thereby avoiding the curse of dimensionality as the system increases in size. As will be shown later, the RL control elements can in fact be effectively trained using only localised inputs. The determination of the reinforcement signal r(t) for the sample system is then based on the performance of the two modelled subsystems (Reservoirs 1 and 2) as well as all sibling subsystems contained in the water supply system. This process is shown in Figure 3.

Case Study

The Reinforcement Learning (RL) model was developed and trained for the Burncrooks reservoir complex located in the Kilpatrick Hills, north of the City of Glasgow, Scotland[4]. This set of five reservoirs,



(a) Input signal to control element

(b) Response of control element

Figure 2: Sample subsystem with input signals *(a)* and output signals *(b)*

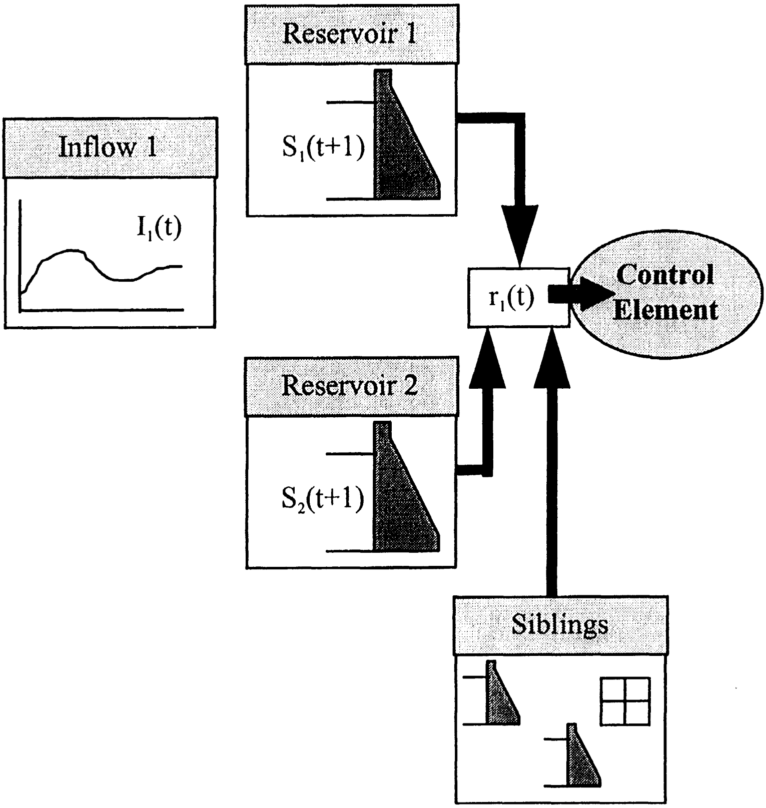


Figure 3: Sources of reinforcement signals

established as a conjunctive use system in 1995, supplies water to the Burncrooks water treatment works (WTW). The system consists of a pumped main transferring water from the Greenside reservoir to the joint Cochno-Jaw reservoirs. The Cochno and Jaw reservoirs are two reservoirs separated by an embankment, and are operated as a single entity with flows from Cochno to Jaw occurring through gravity. Water from the Jaw reservoir is then pumped to the Kilmannan reservoir before being transferred via a pumped main to the Burncrooks reservoir. A gravity main conveys water from the Burncrooks reservoir to the Burncrooks WTW.

The water supply system was modelled using three loosely connected RL control elements, each representing the control of one of the three pumped main subsystems. A schematic of this system is

Transactions on Ecology and the Environment vol 19, © 1998 WIT Press, www.witpress.com, ISSN 1743-3541

provided in Figure 4. The links in Figure 4 relate to the state inputs to the RL control elements. Each control element receives only a localised set of state variables, and uses this limited representation of the full supply system to select one of four possible pump settings [0%, 33%, 66% and 100% of pump capacity]. The control elements are trained to eliminate the following failure events:

(a) Failure to meet water demand at Burncrooks WTW;
(b) Reservoir level below the specified minimum level;
(c) Reservoir spilling while excess storage capacity remains in the system;
(d) Water transfer when downstream reservoir is full.

Failure event *(a)* is the only global failure considered in this case study, with the remaining 3 events reflect failures at a local level (local to the control element).

Simple continuity equations were developed for each reservoir in the system to ensure conservation of mass within the system. Due to a lack of streamflow records, historical net inflow records were reconstructed from historical storage levels and water transfer records.

The control elements were trained using 3 years of reconstructed historical net inflows, and training was performed until all failure events had been eliminated. Full training was achieved after 200 iterations of the net inflow sequence. It is recognised that a 3 year net inflow sequence is unlikely to contain the critical period for the system, thereby producing control strategies that may be inappropriate for drought conditions. Current research activities include investigations into the selection of inflow sequences to optimise the training of RL models.

The RL control system trained on the 3 year sequence was tested on a fourth year of net inflow records. The resulting performance of the reservoir complex in this fourth year was good, with only a single system failure. While such a failure rate might be considered too high, a closer inspection of the failure event revealed that it occurred in the first few weekly time steps of the simulation, where the initial storage level selected was so low that no control strategy could have avoided the resulting failure. Despite the fact that once again the critical period for the system was not contained within the sequence of inflows used in the testing, the results are encouraging and tend to suggest that the control strategies developed during training are appropriate for different sequences of flows.

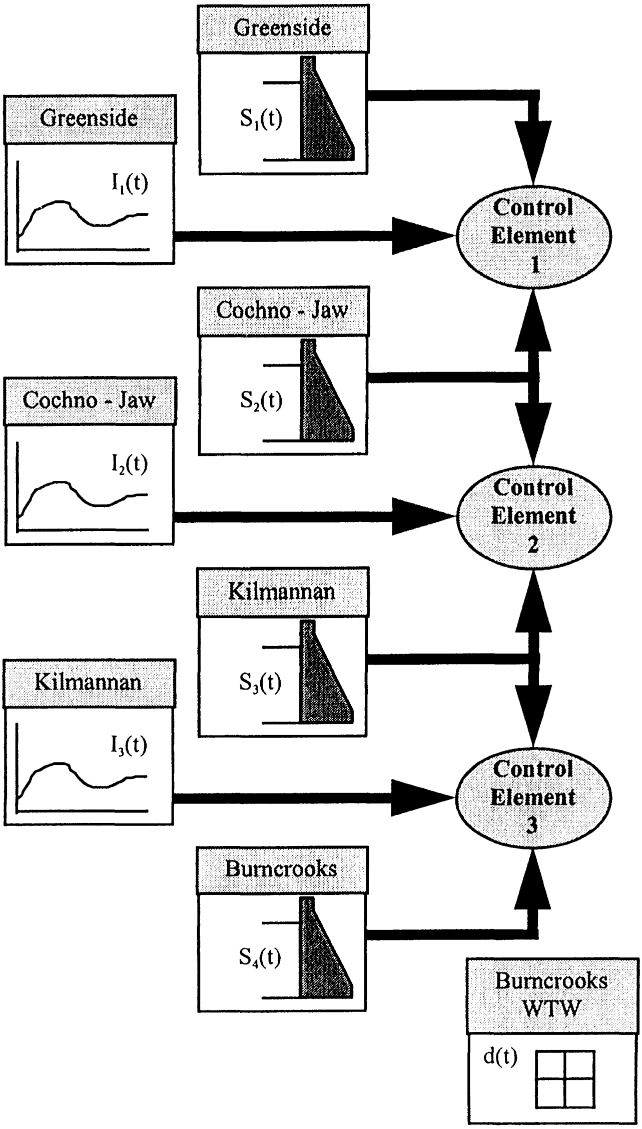


Figure 4: RL schematic of Burncrooks reservoir complex

Conclusions

A methodology capable of addressing the discrete nature of system failures has been developed based on Reinforcement Learning theory. The resulting formulation provides on-line training of operational control elements with the aim of eliminating system failures. Operational control is localised to subsystems within the complete water supply system. Operational decisions for a given control element are based on the local state conditions of the subsystem being controlled and on a combination of local and global performance indicators. The strength of such an approach is the avoidance of the *curse of dimensionality* which affects many analytical techniques such as Dynamic Programming. Each controller does not require any information regarding the states of the subsystems outside its control. An application of the RL methodology to the Burncrooks reservoir complex in Scotland reveal that this distributed control paradigm does converge to a feasible operational control strategy, thereby yielding operational decisions which avoid the realisation of failure events.

References

[1] Yeh, W.W.-G., Reservoir management and operations models: A state-of-the-art review, *Water Resources Research*, **21(12)**, pp. 1797-1818, 1985.

[2] Barto, A.G., Sutton, R.S. & Anderson, C.W., Neuronlike adaptive elements that can solve difficult learning control problems, *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-13(5)**, pp. 834-846, 1983.

[3] Cohen, P.R. & Feigenbaum, E.A., *The Handbook of Artificial Intelligence*, Vol. 3, Kauffman, Los Altos, Cal., 1982.

[4] Chkam, H., A decision support system for the operation of the Burncrooks reservoir complex, *MSc Thesis*, Heriot-Watt University, 1997.